IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of)
)
KIMURA et al.)
)
Application Number: To be assigned)
)
Filed: Concurrently herewith)
)
For: METHOD FOR INDICATING RELATIONSHIP BETWEEN cDNA SEQUENCE AND GENOME RECORDING MEDIUM, SEQUENCER APPARATUS, AND METHOD FOR DESIGNING A PRIMER)

Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION

Sir:

In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of September 25, 2000, the filing date of the corresponding Japanese patent application 2000-289728.

The certified copy of corresponding Japanese patent application 2000-289728 is submitted herewith. Acknowledgment of receipt of the certified copy is respectfully requested in due course.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344

REED SMITH HAZEL & THOMAS LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
August 21, 2001

日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２０００年　９月２５日

出願番号
Application Number: 特願２０００－２８９７２８

出願人
Applicant(s): 株式会社日立製作所

２００１年　６月　１日

特許庁長官
Commissioner,
Japan Patent Office 及川耕造

出証番号　出証特２００１－３０５１６５６

【書類名】　特許願

【整理番号】　H001612

【提出日】　平成12年 9月25日

【あて先】　特許庁長官　殿

【国際特許分類】　G01N 33/50

【発明者】

【住所又は居所】　東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内

【氏名】　木村　宏一

【発明者】

【住所又は居所】　東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所　中央研究所内

【氏名】　西川　哲夫

【特許出願人】

【識別番号】　000005108

【氏名又は名称】　株式会社　日立製作所

【代理人】

【識別番号】　100091096

【弁理士】

【氏名又は名称】　平木　祐輔

【手数料の表示】

【予納台帳番号】　015244

【納付金額】　21,000円

【その他】　国等の委託研究の成果に係る特許出願（平成１２年度新エネルギー・産業技術総合開発機構（再）委託研究、産業活力再生特別措置法第３０条の適用を受けるもの）

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

出証特２００１－３０５１６５６

【物件名】　　要約書　1

【プルーフの要否】　　要

【書類名】　明細書

【発明の名称】　ｃＤＮＡ配列とゲノム配列との対応表示方法、記録媒体、シーケンサ装置及びプライマ設計方法

【特許請求の範囲】

【請求項１】　グラフの１の軸にゲノム配列上の塩基位置を、他の軸にｃＤＮＡ配列上の塩基位置をとり、前記ゲノム配列のうち所定塩基長以上を有する部分配列について、前記ｃＤＮＡ配列との間で所定の割合以上の類似性を有する部分をグラフ上に線分で表示することを特徴とするｃＤＮＡ配列とゲノム配列との対応表示方法。

【請求項２】　複数のｃＤＮＡを縦軸にとり、前記ｃＤＮＡとの対応関係をｃＤＮＡごとに異なる色で表示することを特徴とする請求項１記載のｃＤＮＡ配列とゲノム配列との対応表示方法。

【請求項３】　ゲノム配列とｃＤＮＡ配列とを入力するステップと、前記ゲノム配列中の所定の塩基長以上を有する部分配列について前記ｃＤＮＡ配列との間で所定の割合以上の類似性を有する部分を検索するステップと、前記ゲノム配列とｃＤＮＡ配列をそれぞれグラフの縦軸と横軸又は横軸と縦軸にとって前記検索するステップで検索された部分を線分で該グラフ上に表示するステップと、を備えるｃＤＮＡ配列とゲノム配列との対応表示方法をコンピュータに実行させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項４】　さらに前記所定の塩基長及び前記類似性の所定の割合を入力するステップを備えるｃＤＮＡ配列とゲノム配列との対応表示方法をコンピュータに実行させるためのプログラムを記録したことを特徴とする請求項３記載の記録媒体。

【請求項５】　ネットワーク接続された又は内蔵されたゲノムデータベースにアクセスし、ゲノム配列を入力し、シーケンシングによって得られたｃＤＮＡ配列を入力する入力手段と、前記ゲノム配列中の所定の塩基長以上を有する部分配列について前記ｃＤＮＡ配列との間で所定の割合以上の類似性を有する部分を検索する検索手段と、前記ゲノム配列とｃＤＮＡ配列をそれぞれグラフの縦軸と

横軸又は横軸と縦軸にとって前記検索手段で検索された部分を線分で該グラフ上に表示して前記ｃＤＮＡ配列に対応する前記ゲノム配列上の遺伝子のエクソン・イントロン構造を表示する表示手段とを備えることを特徴とするシーケンサ装置。

【請求項６】　イントロン配列を跨ぐ相異なるエクソン領域内にあるプライマのペアを設計しこれを用いてゲノムライブラリとｃＤＮＡライブラリとでそれぞれＰＣＲを行うステップと、該ＰＣＲを行うステップで増幅されたゲノム配列とｃＤＮＡ配列とを入力するステップと、前記ゲノム配列中の所定の塩基長以上を有する部分配列について前記ｃＤＮＡ配列との間で所定の割合以上の類似性を有する部分を検索するステップと、前記ゲノム配列とｃＤＮＡ配列をそれぞれグラフの縦軸と横軸又は横軸と縦軸にとって前記検索するステップで検索された部分を線分で該グラフ上に表示することによってイントロン配列が存在するために異なったポリヌクレオチドが増幅されたことを表示して増幅したゲノム配列がイントロン配列を含んでいることを確認するステップと、を備えることを特徴とするプライマ設計方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は遺伝子配列の情報解析に係わり、ｃＤＮＡとゲノムの配列類似性検索結果から、ゲノム上の遺伝子の位置と構造を推定し表示する方法に関する。

【０００２】

【従来の技術】

ゲノム上の遺伝子の位置とそのエクソン・イントロン構造を推定する方法としては、ｃＤＮＡ配列とゲノム配列との類似性検索を行い、類似性のある部分配列区間を列挙する方法がある。このとき、類似性のある部分配列区間は、類似度が高い順にソートして列挙される。類似度は、そのような類似性が偶然現れる確率によって評価し、その確率値が小さいものほど類似度が高いとする。

【０００３】

このようなソート法が有用である理由は、以下のように考えられる。生物のゲ

ノムは、遺伝子のコピーを派生させ分化させることにより進化してきた。そのため、一般に、一つのｃＤＮＡ配列に対して、ゲノム上の複数箇所に、種々の類似度で類似する部分配列が存在する。それら複数のゲノム部分配列のうち、実際にそのｃＤＮＡの鋳型となったｍＲＮＡに転写されたゲノム部分配列は、類似度が最も高いものに限られる。このときの不一致部分は、ＳＮＰなどの多型に起因するか、または、シーケンシング・エラーによるものと考えられる。従って、類似性のある区間を類似性の高い順にソートして列挙することにより、そのｃＤＮＡの鋳型となったｍＲＮＡに転写されたゲノム上の部分配列が上位に列挙され、ｃＤＮＡ配列とゲノム配列との対応付けが容易になる。

【０００４】

また、ｃＤＮＡ配列とゲノム配列との対応においては、ｃＤＮＡ配列全体が一本の配列としてゲノム内の部分配列に対応することは少なく、一般には、ｃＤＮＡ配列は何本かの部分配列に分かれ、その各々がゲノム内の部分配列に対応する。このような対応が見られる理由は、ヒトを含む真核生物において、ゲノムからｍＲＮＡが合成される際、スプライシングと呼ばれる現象が起きることによる。ｃＤＮＡとゲノム上で対応している各々の部分配列はエクソンとよばれる。ｃＤＮＡ上ではエクソンは切れ目なく繋がっているが、ゲノム上ではイントロンとよばれる部分配列を挟んで繋がっている。ｃＤＮＡ上のエクソンとゲノム上のエクソンの位置関係は次のいずれかになっている。

【０００５】

（１）ｃＤＮＡ上の各エクソン配列とゲノム上の各エクソン配列はほぼ一致し（以下、これらは向きが同じという）、それらは同じ順番で並んでいる。

（２）ｃＤＮＡ上の各エクソン配列とゲノム上の各エクソン配列は互いにほぼ相補鎖の関係にあり（以下、これらは向きが逆という）、それらは互いに反対の順番で並んでいる。

【０００６】

このようなエクソン・イントロン構造をもつｃＤＮＡ配列とゲノム配列との対応の様子は、類似性のある区間の列挙だけでは把握できず、それら類似性のある区間の相互の位置を調べる必要がある。そのためには、ゲノム配列上の塩基位置

とｃＤＮＡ配列上の塩基位置を両軸にとった２次元プロットが役立つ。最も単純なプロット法として、ゲノム配列のｘ塩基目とｃＤＮＡ配列のｙ塩基目が同一の塩基であるとき、２次元上の座標（ｘ，ｙ）に点をプロットする方法（ドットマトリクス法）がある（105頁、Sequence Analysis Primer, M. Gribskov and J. Devereux, Oxford University Press, 1992年）。この方法では、局所的に精緻な比較が可能となる。また、より大局的な対応関係を捉える方法として、ゲノム配列内とｃＤＮＡ配列内に一定塩基長のウィンドウをとり、これらのウィンドウ内の塩基配列が一定割合以上類似しているとき、ゲノム配列内のウィンドウ位置をｘ軸にｃＤＮＡ配列内のウィンドウ位置をy 軸にとり、それらのウィンドウに対応する線分を２次元平面上にプロットする方法がある（108頁、Sequence Analysis Primer, M. Gribskov and J. Devereux, Oxford University Press, 1992年）。この方法では、一塩基ずつの比較ではなく、数塩基～数十塩基ずつの平均的な比較が行われるため、より長い配列同士の比較が可能になり、また、偶然生じ意味をもたない短い一致部分を排除できる。

【０００７】

【発明が解決しようとする課題】

エクソン・イントロン構造をもつｃＤＮＡ配列とゲノム配列との対応関係を、判り易くグラフィック表示する。ゲノム上には多数の遺伝子が存在する領域があり、多数のｃＤＮＡが対応付けられる（貼り付けられるとも言う）ことがあり、それらの位置関係はグラフィック表示することにより、視覚的に理解しやすくなる。

【０００８】

また、遺伝子のエクソン・イントロン構造において、イントロン配列はエクソン配列に比較して極めて長いことがある。ｃＤＮＡ配列の長さは概ね数百から数万塩基長程度であるが、ゲノム上の対応する遺伝子領域は百万塩基長のオーダーまで広がることがある。このようにｃＤＮＡとゲノムとで対応させるべき配列の長さが３桁も異なる場合には、同じサイズのウィンドウをｃＤＮＡ配列内とゲノム配列内で移動して調べる従来の方法は非効率的となる。

【０００９】

また、ゲノム上の広い範囲にわたってｃＤＮＡとの類似配列の位置を表示する場合、真の対応関係に関与しない多数の類似配列が現われ、真の対応関係を２次元表示の中から拾い出すことを妨げる。そのようなものとして、短い類似配列や、類似度の低い類似配列、向きや順番が不整合の類似配列などが考えられる。そこで、これらの不要な類似配列を除去することが必要になる。

【００１０】

【課題を解決するための手段】

本発明では、与えられたｃＤＮＡ配列とゲノム断片配列に対して、以下の処理ステップから構成される方法によって、それらの間のエクソン・イントロン構造をもった対応関係を表示する。

（１）与えられたｃＤＮＡ配列を纏めて検索用にデータベース化しておき、与えられた各ゲノム断片配列ごとに、それを検索配列としてｃＤＮＡ配列データベースに対して類似性検索を繰り返し行うステップ。

【００１１】

（２）互いに類似性があるｃＤＮＡとゲノムの部分配列のペアを列挙し、そのペアの特徴量として、①部分配列の塩基長、②類似度、③各部分配列がゲノム配列上またはｃＤＮＡ配列上で並ぶ向きと順番、④ｃＤＮＡ部分配列が他のペアのｃＤＮＡ部分配列と共同してｃＤＮＡ配列全体を被覆できる割合、を計算するステップ。

【００１２】

（３）前項で列挙された類似性のある部分配列ペアの集合の中から、上記の特徴量に関する所定の緩い条件を満たさないものを削除するステップ。これは、意味のある類似性を反映している可能性が低いものを除去して処理量を圧縮することを目的とする。即ち、所定の長さや所定の類似度に満たないもの、また、ゲノム上で互いに整合性のある向きと順番をとりえないもの、また、共同してｃＤＮＡ配列の所定以上の割合を覆う可能性のないものを除去する。

【００１３】

（４）前ステップで選び出された類似性のある部分配列ペアの集合の中から、上記の特徴量に関して更に厳格な条件により、表示すべきペアの集合を絞り込むス

テップ。これは、意味のある類似性を反映している可能性が高いものを正確に選び出すことを目的とする。そのためには、例えば、グラフィック表示を利用し、ユーザからの対話的な指示により絞込みの条件の閾値を与えるパラメータを調整する。または、ゲノム上に互いに整合性のある向きと順番で現われ、共同してｃＤＮＡ配列の所定以上の割合を覆うことができる部分配列の集合を、プログラムに従って自動的に選び出し、結果をグラフィック表示する。

【００１４】

（５）選び出されたｃＤＮＡとゲノムの部分配列ペアの位置関係を２次元的に表示するステップ。グラフの１の軸にゲノム配列上の塩基位置を、他の軸にｃＤＮＡ配列上の塩基位置をとり、各部分配列ペアを一本の線分で表示する。この線分は、それぞれ軸へ射影したときに部分配列の位置を表し、かつ、ｃＤＮＡとゲノムの向きの対応を表す。

【００１５】

このため、本発明のｃＤＮＡ配列とゲノム配列との対応表示方法は、グラフの１の軸にゲノム配列上の塩基位置を、他の軸にｃＤＮＡ配列上の塩基位置をとり、前記ゲノム配列のうち所定塩基長以上を有する部分配列について、前記ｃＤＮＡ配列との間で所定の割合以上の類似性を有する部分をグラフ上に線分で表示することを特徴とする。

また、複数のｃＤＮＡを縦軸にとり、前記ｃＤＮＡとの対応関係をｃＤＮＡごとに異なる色で表示することが好ましい。

【００１６】

また、本発明は、ゲノム配列とｃＤＮＡ配列とを入力するステップと、前記ゲノム配列中の所定の塩基長以上を有する部分配列について前記ｃＤＮＡ配列との間で所定の割合以上の類似性を有する部分を検索するステップと、前記ゲノム配列とｃＤＮＡ配列をそれぞれグラフの縦軸と横軸又は横軸と縦軸にとって前記検索するステップで検索された部分を線分で該グラフ上に表示するステップと、を備えるｃＤＮＡ配列とゲノム配列との対応表示方法をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体である。

【００１７】

さらにｃＤＮＡ配列とゲノム配列との対応表示方法は、前記所定の塩基長及び前記類似性の所定の割合を入力するステップを備えることが好ましい。

また、本発明のシーケンサ装置は、ネットワーク接続された又は内蔵されたゲノムデータベースにアクセスし、ゲノム配列を入力し、シーケンシングによって得られたｃＤＮＡ配列を入力する入力手段と、前記ゲノム配列中の所定の塩基長以上を有する部分配列について前記ｃＤＮＡ配列との間で所定の割合以上の類似性を有する部分を検索する検索手段と、前記ゲノム配列とｃＤＮＡ配列をそれぞれグラフの縦軸と横軸又は横軸と縦軸にとって前記検索手段で検索された部分を線分で該グラフ上に表示して前記ｃＤＮＡ配列に対応する前記ゲノム配列上の遺伝子のエクソン・イントロン構造を表示する表示手段とを備えることを特徴とする。

【００１８】

また、本発明のプライマ設計方法は、イントロン配列を跨ぐ相異なるエクソン領域内にあるプライマのペアを設計しこれを用いてゲノムライブラリとｃＤＮＡライブラリとでそれぞれＰＣＲを行うステップと、該ＰＣＲを行うステップで増幅されたゲノム配列とｃＤＮＡ配列とを入力するステップと、前記ゲノム配列中の所定の塩基長以上を有する部分配列について前記ｃＤＮＡ配列との間で所定の割合以上の類似性を有する部分を検索するステップと、前記ゲノム配列とｃＤＮＡ配列をそれぞれグラフの縦軸と横軸又は横軸と縦軸にとって前記検索するステップで検索された部分を線分で該グラフ上に表示することによってイントロン配列が存在するために異なったポリヌクレオチドが増幅されたことを表示して増幅したゲノム配列がイントロン配列を含んでいることを確認するステップと、を備えることを特徴とする。

【００１９】

【発明の実施の形態】

以下、本発明の実施の形態を、図を用いて詳細に説明する。

図１に、与えられたｃＤＮＡ配列をデータベース内のゲノム配列に貼り付けることにより、ｃＤＮＡに対応する遺伝子のエクソン・イントロン構造を可視化することを目的とした、本発明の一実施例における処理の流れを示す。

【００２０】

図１において、101は解析の対象とするｃＤＮＡ配列データであり、102はｃＤＮＡ配列と比較されるべきゲノム配列を格納したデータベースである。103は、ｃＤＮＡ配列データとゲノム配列データベースを読み込む入力処理である。104は、以後の類似性検索に備えるために、入力されたｃＤＮＡ配列データをデータベース化する処理であり、公知の方法を用いたプログラムformatdb (Altschul,Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.）を使う。105は、ゲノムデータベース内にある各々のゲノム断片配列ごとに、それを検索配列としてｃＤＮＡデータベースに対して類似性検索処理を繰り返す処理である。この各々の類似性検索処理は、公知のアルゴリズムを用いたプログラムであるBLAST (Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.)を用いる。106は、各ゲノム断片配列ごとに得られた類似性検索結果を記述したテキストデータを全て読み込み、その中に現われる類似性がある部分配列を抽出して列挙し、その各々の部分配列を特徴付ける諸量を計算する処理である。107は、それらの諸特徴量に基づき、列挙された類似性のある部分配列の中から、所定の緩い条件を満たすものを選び出す、類似部分配列の１次選択処理である。これは、意味のある類似性を反映している可能性が低いものを除去して処理量を圧縮することを目的とする。その選択結果をファイル108に記憶する。ここまでの計算処理は時間を要するため、また、ここまでの計算は以後のユーザとの対話的処理とは独立に１回だけ行えばよいため、このようにファイルに記憶しておく。109は、ｃＤＮＡ上とゲノム上にある類似性のある部分配列のうちで選択されたものの相互の位置関係をファイル108から読み込んで、ユーザに分かり易く呈示するために、２次元のグラフィック表示データを生成する処理である。110はモニターディスプレイ，キーボード，マウスを備えたユーザインターフェース装置

で、109で生成されたグラフィックデータを表示するとともに、ユーザからの描画パラメータを受け付けて109に渡してグラフィックデータを再計算させ、109と110は共同して対話的な表示を行う。さらに、111は、更に厳格な条件により部分配列を更に絞り込む、類似部分配列の２次選択処理である。これは、意味のある類似性を反映している可能性が高いものをより正確に選び出すことを目的とする。110は、そのために必要となるパラメータをユーザから受け付け、それらを111に送る。111により更に絞り込まれた類似部分配列のデータは109に送られ、そこで、グラフィックデータが再計算される。これは、再び、110に送られ、ユーザに呈示される。109と110と111とにより、対話的に部分配列の選択の仕方を対話的に変更することができ、これにより、ゲノムとｃＤＮＡとの対応関係を正しく捉えた部分配列の集合を選び出すことができる。

【００２１】

図２は、106において、類似性のあるゲノム断片配列の部分配列とｃＤＮＡ部分配列とのペアを抽出して得られるデータのデータ構造を表す。ここに現われる情報は全て、105のBLAST プログラムによる類似性検索結果の中から得ることができる。201は、１本のゲノム断片配列に対応するデータであり、全体のデータはこの繰り返し構造をもつ。201は、少なくとも、ゲノム断片配列を識別する名前とその配列長、及び、そのゲノム断片配列と類似性のある部分配列を持つｃＤＮＡに関する情報202の繰り返し構造を含む。202は、少なくとも、ｃＤＮＡを識別する名前とその配列長、及び、ゲノムとの類似性のある部分配列に関する情報203の繰り返し構造を含む。以後、説明の簡略化のため、ゲノム内とｃＤＮＡ内にある互いに類似性のある部分配列を“エクソン”と呼ぶことにする。これは生物学的なエクソンに対応することもあるが、それ以外に、偶然生じた類似性による部分配列のペアも含むこともある。203はエクソンの情報であり、少なくとも、長さ、ゲノムとｃＤＮＡとの一致塩基数、ゲノム断片配列内とｃＤＮＡ配列内での位置の情報を含む。

【００２２】

図２に示したデータ構造は、図１内の106以降で処理される情報の基本構造であり、ファイル108に格納される情報もこのデータ構造をもつ。これは、106で得

られた情報から、107において有用性が低いと判断される一部の情報が除去されたものである。109は、図２に示したデータ構造をもつ情報を読み込んでグラフィック表示を行い、また、111は図２に示したデータ構造をもつ情報を読み込んで、そこから有用性が高いと判断されるエクソンを選び出し、再び、図２のデータ構造の情報を109に返す。

【００２３】

図３は、107の類似部分配列ペア（エクソン）の１次選択処理の動作を説明するためのフローチャートである。301の終了判定を含む繰り返し処理により、全てのゲノム断片配列に対して以下の処理を行う。302で、処理中のゲノム断片配列に対する201に示す情報を読み込む。この中には、202に示すｃＤＮＡの情報が複数含まれる。303の終了判定を含む繰り返し処理により、これら全てのｃＤＮＡに対して以下の処理を行う。304で、処理中のｃＤＮＡ配列に対する202に示す情報を読み込む。この中には、203に示すエクソンの情報が複数含まれる。305では、これらの個々のエクソンについて、

（類似度）＝（エクソン内一致塩基数）／（エクソン塩基長）

により類似度を計算し、これが所定の類似度に満たない場合は、203に列挙された中から該当エクソンを削除する。所定の類似度として、例えば80％を設定しておけば、現在処理中のｃＤＮＡの鋳型となった遺伝子（またはその近縁の遺伝子）に含まれるエクソン以外のゲノム断片の部分配列は、ほぼ除去されると考えられる。次に、306では、残ったエクソン長の最大値を求め、それが所定の値以上かどうかを判定する。多くの場合、遺伝子中のエクソンの中には100塩基長程度のもの少なくとも１つはある。したがって、例えば50塩基長程度の長さのエクソンがひとつも見つからないとすれば、この場合、ゲノム中に豊富に遍在する繰り返し配列の一部を捉えている可能性が高いと考えられるので、307によりすべてのエクソン情報とそのｃＤＮＡ情報を除去する。そうでない場合は、エクソン長の合計を計算し、ｃＤＮＡ配列の全長との比を求め、308でその値が所定の値以上かを判定する。その比の値が例えば30％に満たないような場合は、それらのエクソンはｃＤＮＡ配列のごく一部しか覆うことができないため、そこでのｃＤＮＡとゲノムとの関連は薄いと考えられるので、307によりすべてのエクソン情報

とそのｃＤＮＡ情報を除去する。

【００２４】

図４は、109の表示処理により生成され、110のモニター画面上に描画されるイメージを、簡略化して表した説明図である。401は処理したゲノム断片配列のリストであり、その中の１項目（図では「ゲノム断片配列２」）が選択され、その項目に対する解析結果がモニター画面に表示されていることを表している。402は、横軸にゲノム断片配列上の塩基位置を荒い座標系（図ではメガ塩基単位）でとり、縦軸にｃＤＮＡ配列上の塩基位置を細かい座標系（図ではキロ塩基単位）でとり、ゲノムとｃＤＮＡ間の類似部分配列のペアを示すエクソンを線分で表す。これらのエクソンを表す線分は、実際のモニター画面では、ｃＤＮＡごとに色分けして表示する。403は、各ｃＤＮＡに対してエクソンの合併がｃＤＮＡ配列の全体をどの割合まで覆うか示す。これは、そのｃＤＮＡが現在処理中のゲノム断片配列とどの程度強い関連があるかを示している。404はｃＤＮＡ配列のリストであり、その中の１項目（図では「ｃＤＮＡ配列１」）が選択され、その項目に対する解析結果がモニター画面に表示されていることを表している。405は、404において選ばれたｃＤＮＡに対して、それを含む402のプロットの一部を拡大表示したものである。406は、405のエクソンを表す線分のプロットを、縦軸に射影したものである。ここで、エクソンの合併がｃＤＮＡ全体をどの程度覆うかを確認できる。また、407は、405のエクソンを表す線分のプロットを、横軸に射影したものである。ここで、射影されたエクソンに挟まれた部分がイントロンを表す。408は、各エクソンに対して、その塩基長とその中の（ゲノム・ｃＤＮＡ間の）一致塩基数を表示したものである。これにより、各エクソンにおけるゲノム・ｃＤＮＡ間の類似度がどの程度高いかを確認できる。

【００２５】

図５は、111の類似部分配列ペア（エクソン）の２次選択処理の動作を説明するためのフローチャートである。501の終了判定を含む繰り返し処理により、全てのゲノム断片配列に対して以下の処理を行う。502で、処理中のゲノム断片配列に対する201に示す情報を読み込む。この中には、202に示すｃＤＮＡの情報が複数含まれる。503の終了判定を含む繰り返し処理により、これら全てのｃＤＮ

Ａに対して以下の処理を行う。504で、処理中のｃＤＮＡ配列に対する202に示す情報を読み込む。この中には、203に示すエクソンの情報が複数含まれる。505では、これらの個々のエクソンについて、

（類似度）＝（エクソン内一致塩基数）／（エクソン塩基長）

により類似度を計算し、これが所望の類似度に満たない場合は、203に列挙された中から該当エクソンを削除する。所望の類似度は、ユーザインターフェース111によりプログラムに伝えられる。例えば、ここで類似度98％を要求すれば、２％程度の違いはＳＮＰなどの多型またはシーケンシング・エラーによるものと許容して、現在処理中のｃＤＮＡの鋳型となった遺伝子（またはそれに酷似した遺伝子）に含まれるエクソンのみが選ばれると考えられる。次に、506では、残ったエクソンの集合を、向きと順番が互いに整合的であるようなグループに分割する。すなわち、各グループごとに、そこに属するエクソンの集合は次のいずれかの条件を満たす。

【００２６】

（１）ｃＤＮＡ上の各エクソン配列とゲノム上の各エクソン配列はほぼ一致し（これらは向きが同じ、または、正の向きという）、それらは同じ順番で並んでいる。

（２）ｃＤＮＡ上の各エクソン配列とゲノム上の各エクソン配列は互いにほぼ相補鎖の関係にあり（これらは向きが逆、または、負の向きという）、それらは互いに反対の順番で並んでいる。このようなグループ分けを行う手順は後述する。507の終了判定を含む繰り返し処理により、エクソンの各グループに対して以下の処理を行う。508でグループ内に属するエクソンの合併がｃＤＮＡ全体を覆う割合を計算しそれが所定の割合（例えば95％）以上かを判定し、また、グループ内のエクソンをｃＤＮＡ配列上で昇順に並べたとき隣り合うエクソン間の間隔が所定の塩基長（例えば10塩基）未満になっているかを判定し、違反があれば509においてこのグループに属する全エクソンを203から削除する。

【００２７】

１つのｃＤＮＡに属するエクソン全体を、506において上記のようにグループ分けするには、次のような手順に従う。まず、１つのｃＤＮＡに属するエクソン

全体を正・負の向きによって２つに分ける。次に、正の向きのエクソンをゲノム断片配列上の位置により昇順にソートし、また、負の向きのエクソンをゲノム断片配列上の位置により降順にソートする。それぞれの向きのエクソンについてソート順に見ていき、

（１）最初のエクソンは新たなグループに属する。

【００２８】

（２）現在のエクソンｑが直前に見たエクソンｐに対して、

（ｑ右端塩基のｃＤＮＡ配列上での位置）

＞（ｐ右端塩基のｃＤＮＡ配列上での位置）－（許容重なり塩基数）

が成り立つならばｑはｐと同じグループに属し、そうでない場合、ｑは新たなグループに属する。許容重なり塩基数としては、例えば、５塩基程度でよい。

【００２９】

【発明の実施の形態　－　その２】

上記実施例によるｃＤＮＡ配列とゲノム配列との対応表示を利用して、プライマ設計を行うための、本発明の第２の実施形態を、図を用いて詳細に説明する。

一般に、ｃＤＮＡライブラリを作成したとき、そこに含まれるポリヌクレオチドとして、ｃＤＮＡ以外に、その他のゲノムの断片が紛れ込むことがある。従って、ＰＣＲを用いてｃＤＮＡ配列の一部を増幅しようとする際には、それが実際にｃＤＮＡ配列の一部であってそれ以外のゲノム断片ではないことを確認できることが有用である。

上記実施例を利用してプライマを設計することにより、このような確認が可能になる。

【００３０】

図６は、そのようなプライマ設計法を説明する原理図である。601はゲノム上の塩基位置を表す軸であり、602はｃＤＮＡ上の塩基位置を表す軸であり、603と604は一つのｃＤＮＡに属する相異なるエクソンを表す。603と604の塩基配列の中から、公知の方法（田平、林、PCR,PCR-SSCP法、新遺伝子工学ハンドブック、村松・山本編、７５頁、羊土社、１９９９年）によりプライマ配列を選び出す。このプライマ配列のオリゴヌクレオチドを合成して、ｃＤＮＡライブラリに対し

てＰＣＲを行えば、これらのプライマは607、608の位置でｃＤＮＡに結合し、それらに挟まれた609に示すｃＤＮＡの部分配列をもつポリヌクレオチドが増幅される。一方、これと同じプライマを用いて、ゲノムライブラリに対してＰＣＲを行えば、これらのプライマは610、611の位置でゲノムに結合し、それらに挟まれた612に示すゲノムの部分配列をもつポリヌクレオチドが増幅される。このポリヌクレオチドはイントロン配列を含んでいる。従って、これら２種類のＰＣＲで増幅されたポリヌクレオチドの長さは異なる。

【００３１】

これに対して、ｃＤＮＡライブラリの中に紛れ込んだゲノム断片からプライマを設計してしまった場合は、上記のような２種類のＰＣＲで増幅されたポリヌクレオチドは一致する。651はゲノム上の塩基位置を表す軸であり、652はｃＤＮＡ上の塩基位置を表す軸であり、653はエクソンを表す。653の塩基配列の中からプライマ配列を選び出す。このプライマ配列のオリゴヌクレオチドを合成して、ｃＤＮＡライブラリに対してＰＣＲを行えば、これらのプライマは656、657の位置でｃＤＮＡライブラリに含まれるゲノム断片に結合し、それらに挟まれた658に示す部分配列をもつポリヌクレオチドが増幅される。また、これと同じプライマを用いて、ゲノムライブラリに対してＰＣＲを行えば、これらのプライマは659、660の位置でゲノムに結合し、それらに挟まれた661に示す配列をもつポリヌクレオチドが増幅される。これら２種類のＰＣＲで増幅されたポリヌクレオチドは一致する。

【００３２】

このように、同じプライマを用いてｃＤＮＡライブラリとゲノムライブラリに対してＰＣＲで増幅されたポリヌクレオチドの違いを調べることにより，ｃＤＮＡに紛れ込んだゲノム断片ではなくｃＤＮＡの一部を増幅していることが確認できる。

【００３３】

【発明の効果】

エクソン・イントロン構造をもつｃＤＮＡ配列とゲノム配列との対応関係を、向きと順番が整合的な（エクソンに対応する）線分の並びとして、判り易くグラ

フィック表示する。エクソンの候補となる類似部分配列のペアについて、その両端の塩基位置と類似度等をあらかじめ計算しておき、その中からよりエクソンとして確からしい類似部分配列のペアを対話的に選んで描画するため、ゲノム上の広範囲にわたって高速に描画できる。短い類似配列や、類似度の低い類似配列、向きや順番が不整合の類似配列などを自動的に除去して表示するため、ｃＤＮＡ配列とゲノム配列との間の意味のある対応関係のみが描画される。

【図面の簡単な説明】

【図１】

本発明の一実施の形態における処理の流れを示す図。

【図２】

類似部分配列ペア（エクソン）を集めた情報のデータ構造。

【図３】

類似部分配列ペア（エクソン）の１次選択処理の動作を説明するためのフローチャート。

【図４】

モニター画面上に描画されるイメージを、簡略化して表した説明図。

【図５】

類似部分配列ペア（エクソン）の２次選択処理の動作を説明するためのフローチャート。

【図６】

本発明の第２の実施形態におけるプライマ設計法の原理を説明する図

【符号の説明】

101　解析の対象とするｃＤＮＡ配列データ

102　ｃＤＮＡ配列と比較されるべきゲノム配列を格納したデータベース

103　ｃＤＮＡ配列データとゲノム配列データベースを読み込む入力処理

104　類似性検索処理のためにｃＤＮＡ配列データをデータベース化する処理

105　各ゲノム断片配列を検索配列としてｃＤＮＡデータベースに対して類似性検索を繰り返す処理

106　類似性がある部分配列ペア（エクソン）を抽出してその特徴量を計算す

る処理

107　処理量を圧縮することを目的とした、類似部分配列の１次選択処理

108　類似性がある部分配列ペア（エクソン）によりゲノムとｃＤＮＡを対応付けるデータを格納したファイル

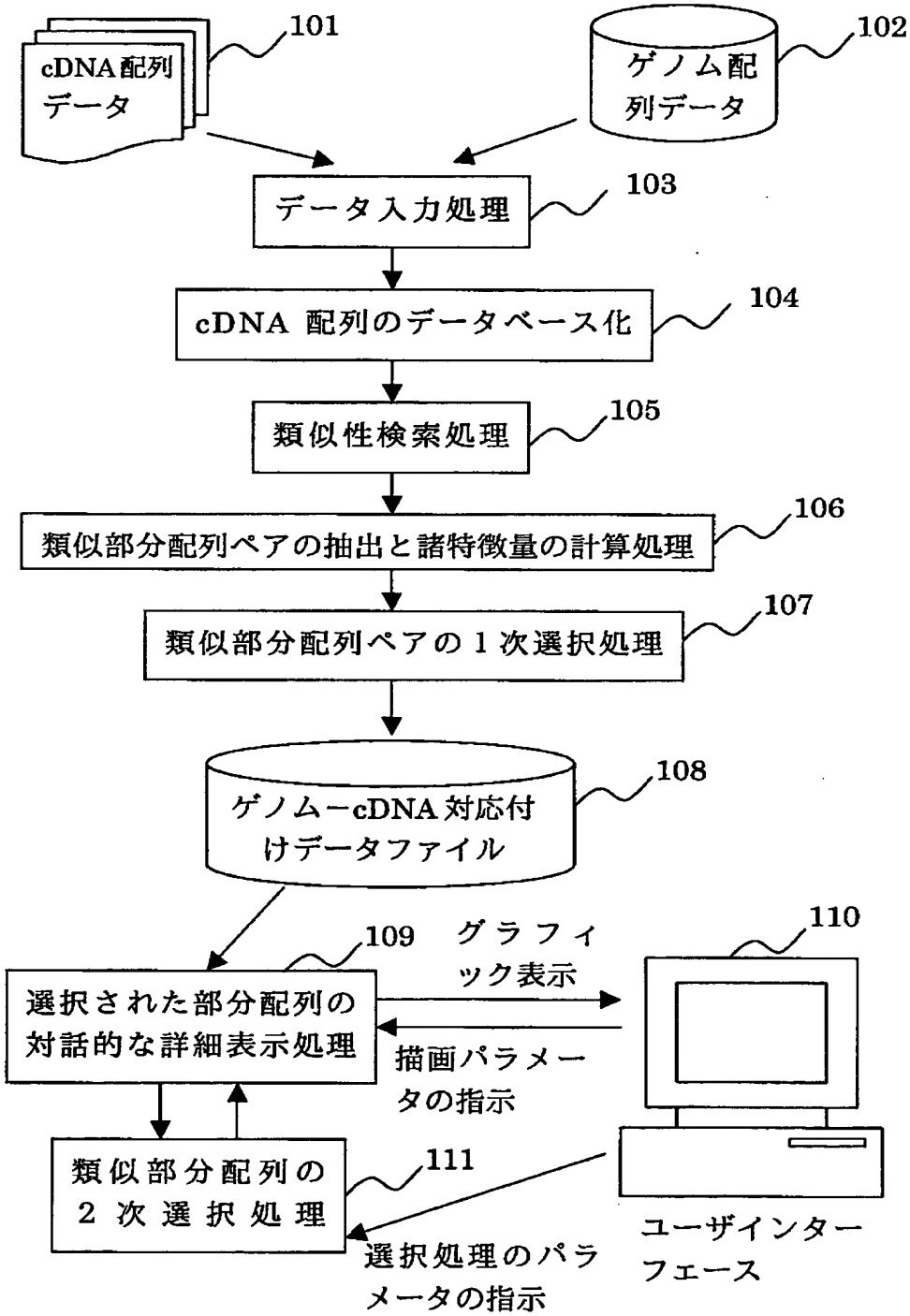
109　２次元のグラフィック表示データを生成する処理

110　ユーザインターフェース装置

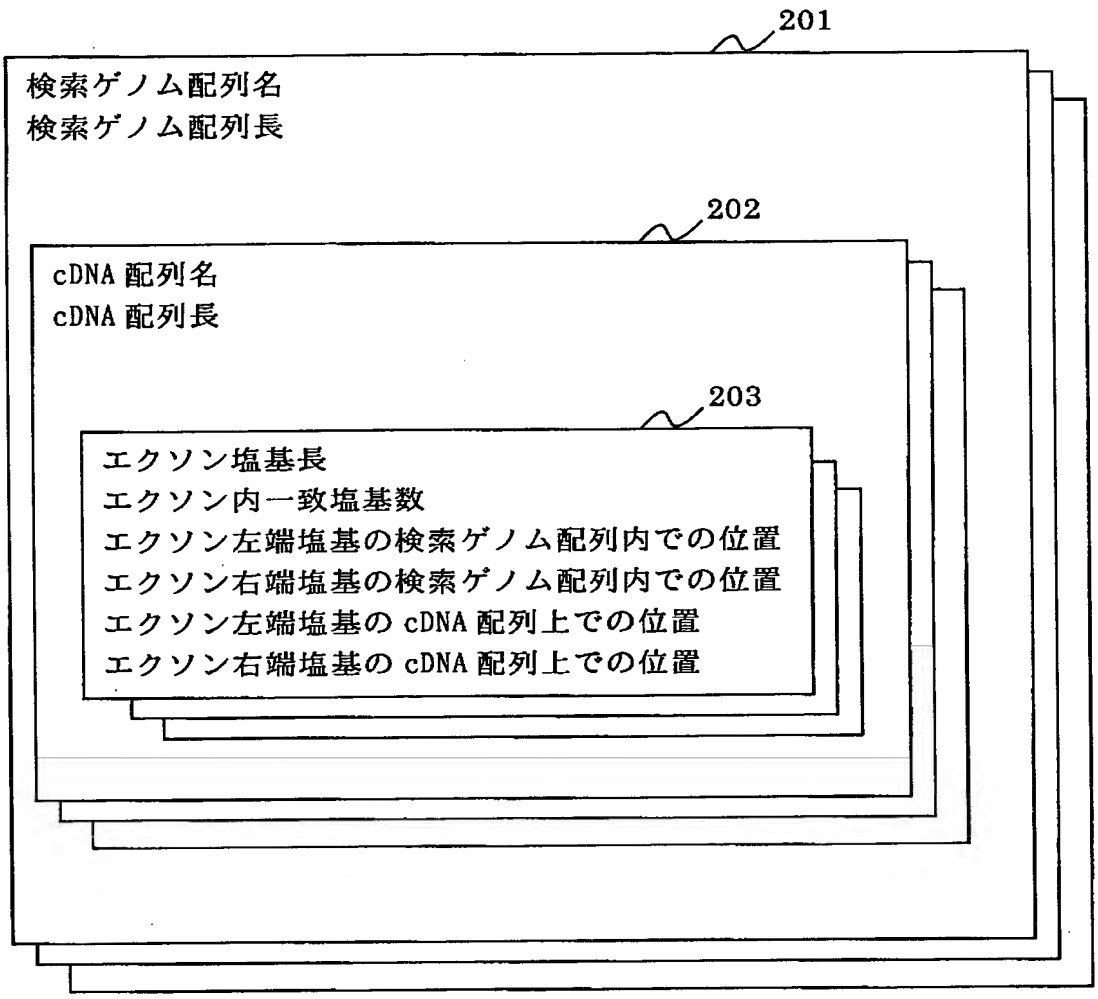
111　意味のある類似部分配列を正確に選び出すための２次選択処理

【書類名】　図面

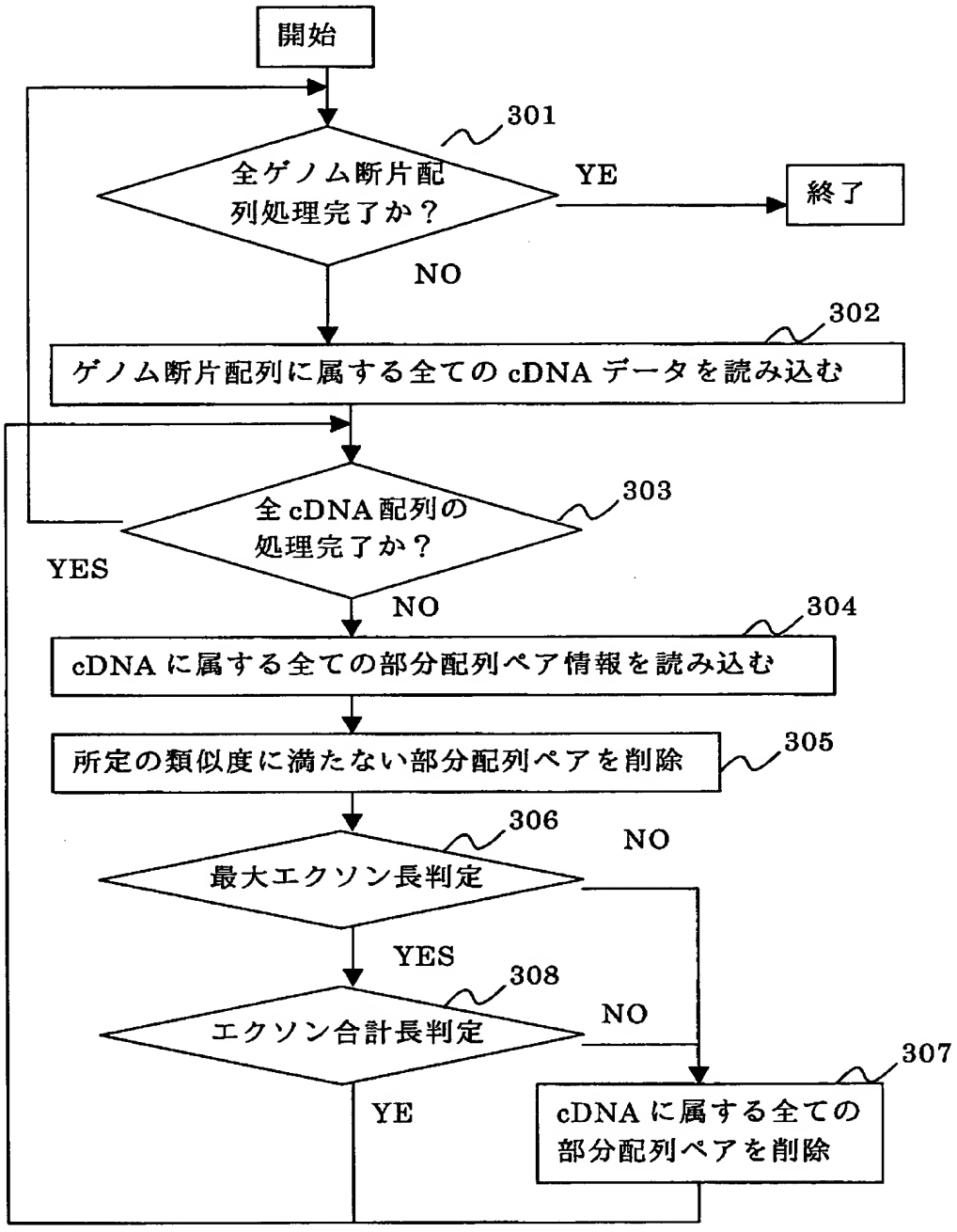
【図１】


101
cDNA配列データ
102
ゲノム配列データ
データ入力処理
103
cDNA 配列のデータベース化
104
類似性検索処理
105
類似部分配列ペアの抽出と諸特徴量の計算処理
106
類似部分配列ペアの１次選択処理
107
ゲノムーcDNA対応付けデータファイル
108
109
選択された部分配列の対話的な詳細表示処理
グラフィック表示
描画パラメータの指示
110
ユーザインターフェース
類似部分配列の２次選択処理
111
選択処理のパラメータの指示

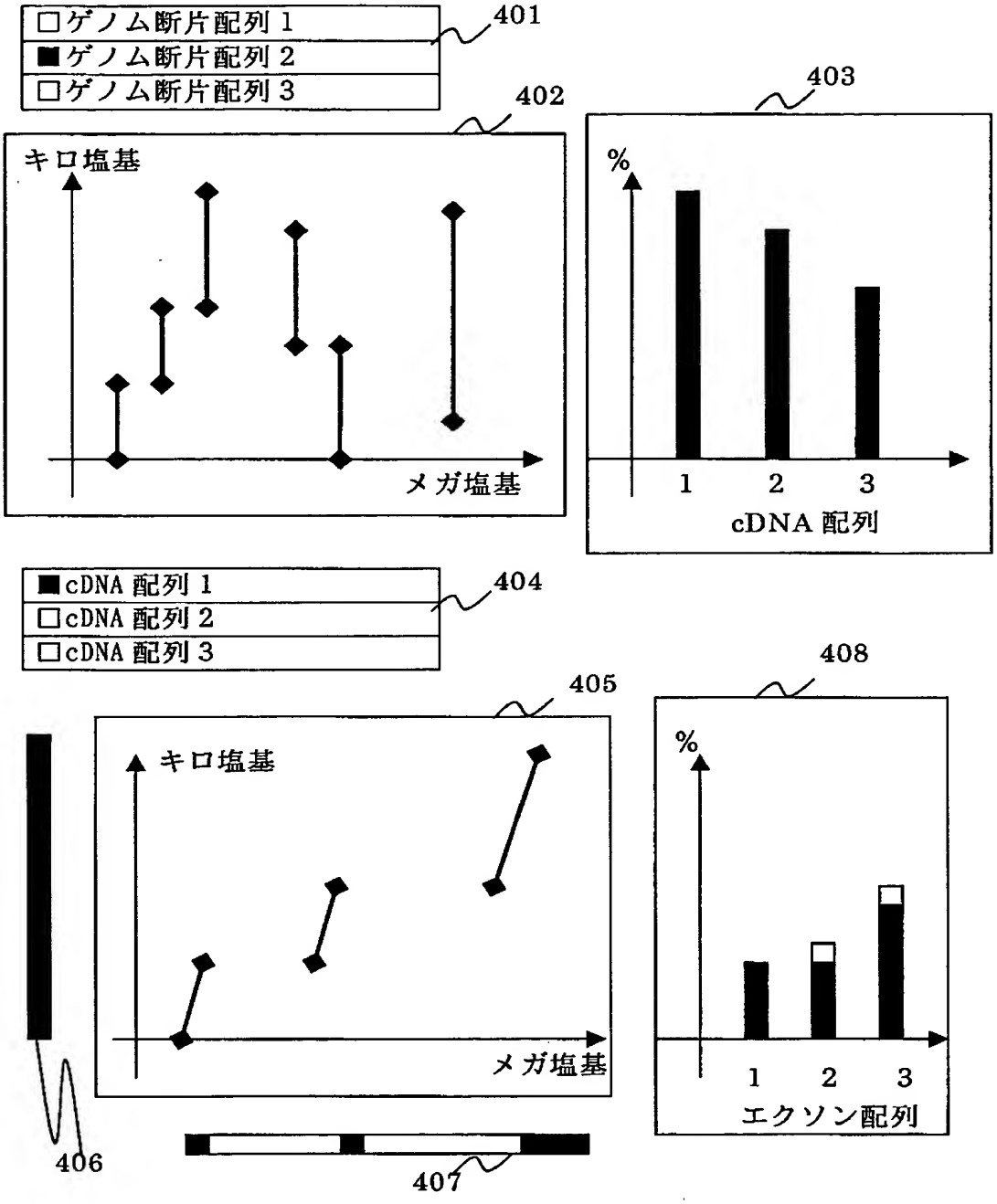
【図２】


201
検索ゲノム配列名
検索ゲノム配列長
202
cDNA 配列名
cDNA 配列長
203
エクソン塩基長
エクソン内一致塩基数
エクソン左端塩基の検索ゲノム配列内での位置
エクソン右端塩基の検索ゲノム配列内での位置
エクソン左端塩基の cDNA 配列上での位置
エクソン右端塩基の cDNA 配列上での位置

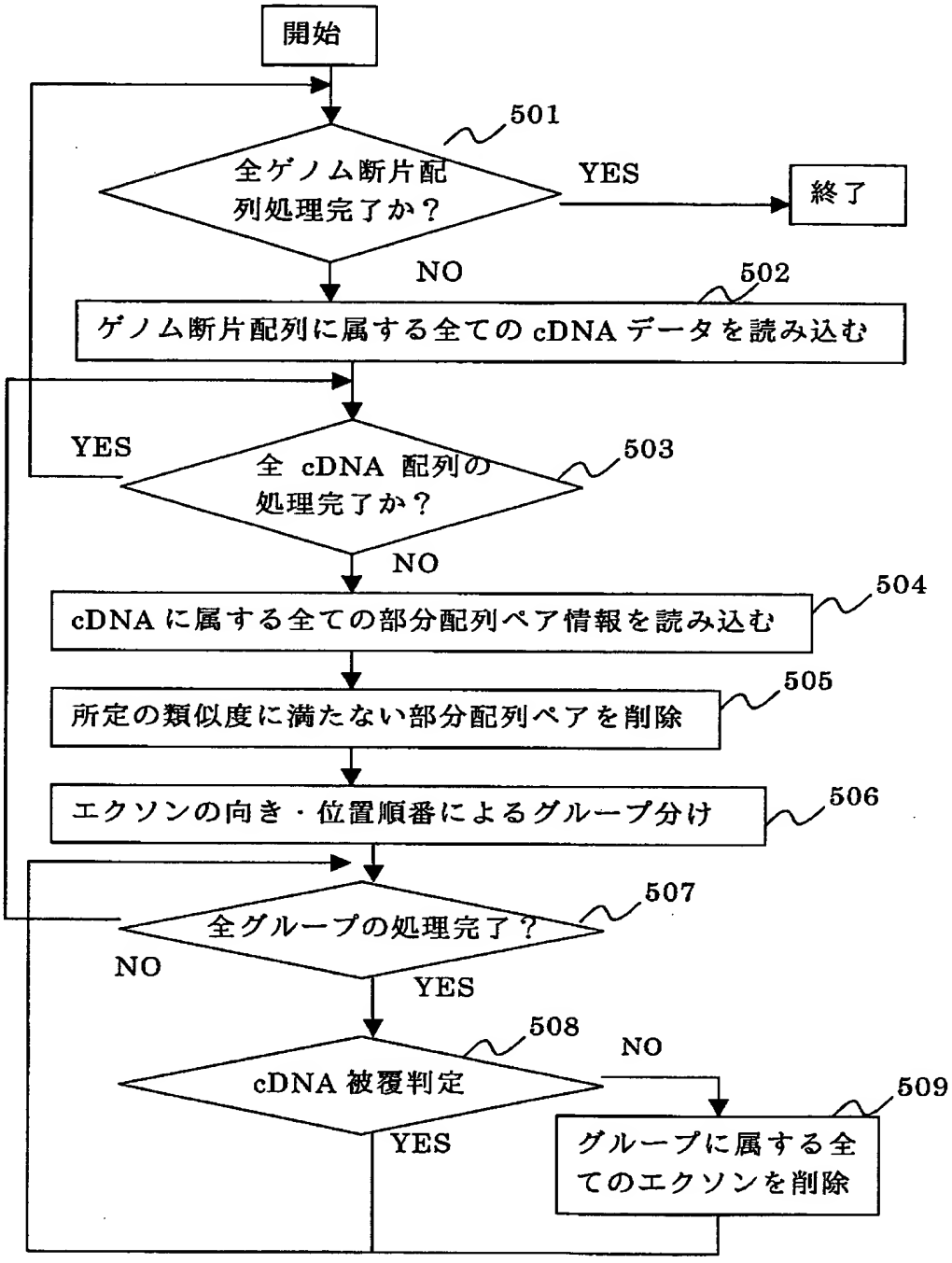
【図３】


開始
301
全ゲノム断片配列処理完了か？
YE
終了
NO
302
ゲノム断片配列に属する全てのcDNAデータを読み込む
303
全cDNA配列の処理完了か？
YES
NO
304
cDNAに属する全ての部分配列ペア情報を読み込む
所定の類似度に満たない部分配列ペアを削除
305
306
最大エクソン長判定
NO
YES
308
エクソン合計長判定
NO
YE
307
cDNAに属する全ての部分配列ペアを削除

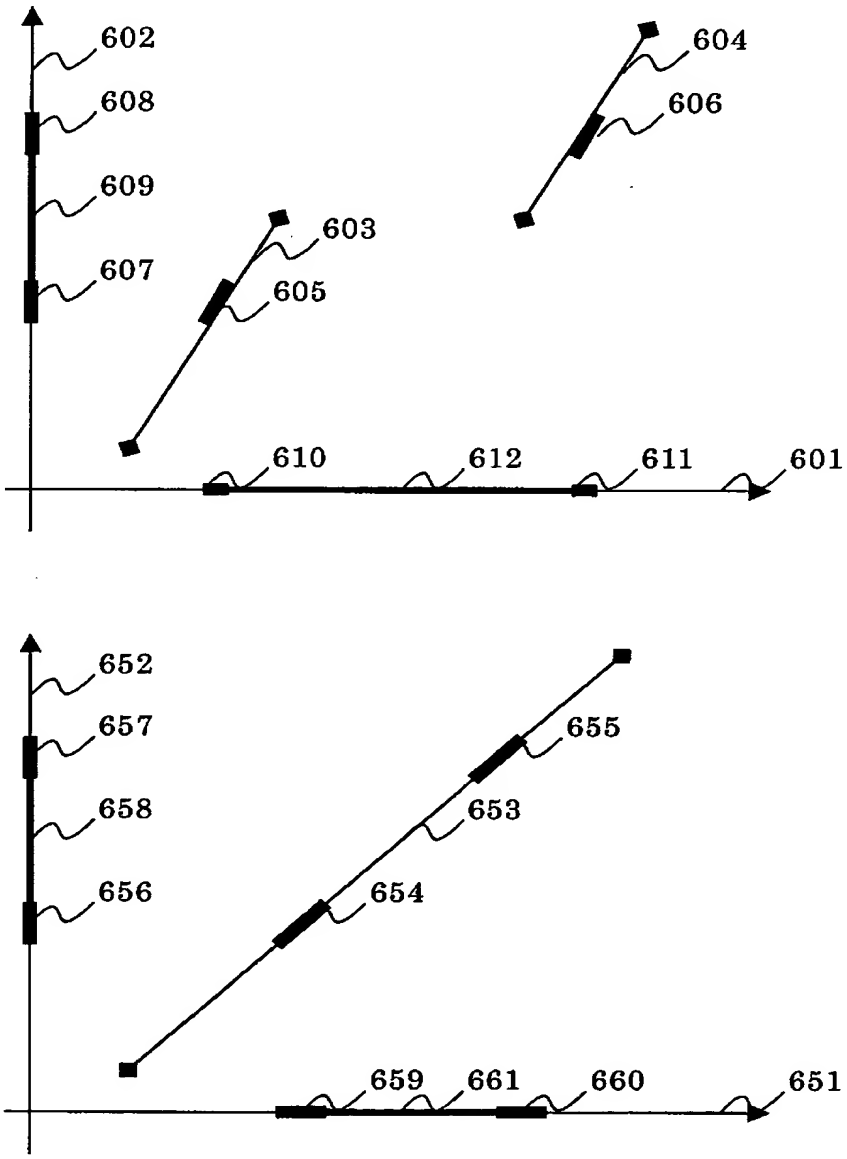
【図４】


□ゲノム断片配列１
■ゲノム断片配列２
□ゲノム断片配列３
401
402
403
キロ塩基
メガ塩基
%
1
2
3
cDNA 配列
■cDNA 配列１
□cDNA 配列２
□cDNA 配列３
404
405
408
キロ塩基
メガ塩基
%
1
2
3
エクソン配列
406
407

【図５】


開始
501
全ゲノム断片配列処理完了か？
YES
終了
NO
502
ゲノム断片配列に属する全ての cDNA データを読み込む
YES
503
全 cDNA 配列の処理完了か？
NO
504
cDNA に属する全ての部分配列ペア情報を読み込む
505
所定の類似度に満たない部分配列ペアを削除
506
エクソンの向き・位置順番によるグループ分け
507
全グループの処理完了？
NO
YES
508
cDNA 被覆判定
NO
509
グループに属する全てのエクソンを削除
YES

【図６】


602
608
609
607
603
605
604
606
610
612
611
601
652
657
658
656
655
653
654
659
661
660
651

【書類名】　　要約書

【要約】

【課題】　エクソン・イントロン構造をもつｃＤＮＡ配列とゲノム配列との対応関係を、判り易くグラフィック表示する表示方法を提供する。

【解決手段】　ｃＤＮＡとゲノムとの類似性検索結果から、類似性のある部分配列のペア（エクソン）の両端の塩基位置・類似度等の情報を抽出する。これらの中から、類似度・配列長などの観点で意味をもつ可能性が低いと判定される部分配列ペアの情報を除去する。また、エクソン間の向き・順番に関する整合性を調べ、ｃＤＮＡを所定の割合以上被覆してｃＤＮＡとの関連が明確になるエクソンのみを選択する。選択されたエクソンを、グラフの１の軸にゲノム配列上の塩基位置を、他の軸にｃＤＮＡ配列上の塩基位置をとり、線分で表示することにより、線分の並びとしてイントロン・エクソン構造が視覚的に確認できる。

【選択図】　図１

出願人履歴情報

識別番号　〔０００００５１０８〕

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所